On Soft-Biometric Information Stored in Biometric Face Embeddings

Philipp Terhörst, Daniel Fährmann, Naser Damer, Florian Kirchbuchner, Arjan Kuijper

Abstract—The success of modern face recognition systems is based on the advances of deeply-learned features. These embeddings aim to encode the identity of an individual such that these can be used for recognition. However, recent works have shown that more information beyond the user's identity is stored in these embeddings, such as demographics, image characteristics, and social traits. This raises privacy and bias concerns in face recognition. We investigate the predictability of 73 different soft-biometric attributes on three popular face embeddings with different learning principles. The experiments were conducted on two publicly available databases. For the evaluation, we trained a massive attribute classifier such that can accurately state the confidence of its predictions. This enables us to derive more sophisticated statements about the attribute predictability. The results demonstrate that the majority of the investigated attributes are encoded in face embeddings. For instance, a strong encoding was found for demographics, haircolors, hairstyles, beards, and accessories. Although face recognition embeddings are trained to be robust against non-permanent factors, we found that specifically these attributes are easily-predictable from face embeddings. We hope our findings will guide future works to develop more privacy-preserving and bias-mitigating face recognition technologies.

Index Terms—Face Recognition, Bias, Fairness, Soft-Biometrics, Analysis, Privacy, Biometrics

1 Introduction

Current face recognition systems show strong recognition capabilities enabled by the advances in learning deep neural feature embeddings [13]. This leads to a worldwide spreading of these systems and also increasingly affect everyone's daily life [8]. Although face recognition models are trained with the aim of extracting deeply-learned features for recognition, recent works showed that the information encoded in such embeddings goes beyond identity. These works showed that different face embeddings contain information about head pose [37], image characteristics (such as quality [4], [18], viewpoint [19], and illumination [35]), demographics [9], [36], [47], and social traits [38]. However, this raises ethical concerns regarding fairness and privacy in face recognition. First, for many applications, the users do only permit to have access to the information related to recognition [32] and extracting additional information without a person's consent is considered a violation of their privacy [24]. This is known as soft-biometric privacy [32] and solutions are either build on image- [30], [31], [34] or embedding-level [5], [42], [45], [51]. Second, the attributes stored in biometric face embeddings can indicate biased performances related to these attributes that might result in unfair performance differences. This is known as face recognition bias and solutions for this problem mainly focused on demographic-bias [12], [28], [48], [52], [55]. To develop more advanced bias-mitigating solutions, knowledge about encoded attributes in face embeddings is required [43].

Extending the work of [43], we provide a predictability analysis of 73 different soft-biometric attributes from face embeddings. In [43], a predictability statement of an attribute is derived by taking into account the attribute prediction performance of an attribute classifier at two difficulty-levels. These difficulty-levels describe how well this classifier can predict an attribute and thus, simulate e.g. different capturing conditions. In contrast to [43], this work

1) analyses the predictability of an attribute in a continuous range of difficulty-levels. This allows deriving more fine-grained predictability statements about single attributes,
2) jointly analyses the predictability of multiple attribute (attribute categories) to extract exact, compact, and easily-understandable findings,
3) visualizes the findings on the attribute predictability to provide the reader with an intuitive understanding of which attributes are stored in face embeddings and how easily these can be predicted,
4) extends the experiments to three different face recognition models. This allows to explore the effect of embedding dimensionalities and the underlying training losses on the attributes encoded in face embeddings.
5) and discusses the implications of our findings on future works.

The investigation methodology is based on a massive attribute classifier (MAC) that is simultaneously trained on multiple attributes to take advantage of a shared feature space. The MAC is constructed such that it can accurately state its prediction confidence [47]. This allows us to derive more detailed statements about the predictability of attributes in face embeddings. The experiments were conducted on two publicly available databases, CelebA [29] and LFW [20], and on three popular face embeddings, FaceNet [40], CosFace [53], and ArcFace [10]. To derive understandable statements about the stored attribute information, we categorized each attribute into one of three

• *All authors are with the Fraunhofer Institute for Computer Graphics Research IGD in Darmstadt (Germany) and with the Technical University of Darmstadt in Darmstadt (Germany)*
E-mail: firstname.lastname@igd.fraunhofer.de

predictability classes: easily-predictable, predictable, and hardly-predictable.

The results demonstrate that many attributes are encoded in face embeddings. From the 113 analysed attributes, 39 attributes are assigned to the easily-predictable class and 74 are predictable. We found differences in the attribute predictability regarding the underlying training principles of the face recognition networks. However, information about age, hairstyles, haircolors, beards, and accessories are strongly encoded in all embedding types, FaceNet, CosFace, and ArcFace. Despite that face embeddings are learned to be robust to non-permanent factors, the results demonstrate that especially these attributes are easily-predictable.

2 Related work

Face recognition benefits from the development of deep neural network representations [13]. However, there is a need to better understand which kind of information is stored in these representations, since these representations are derived from black-box models.

In 2017, Parde et al. [37] demonstrated that the investigated representations contain accurate information about the head position (i.e. the yaw and pitch of a face) and the source of the image (i.e. whether the input-face origins from a still image or a video frame). They suggested that information about the image-quality might also be available in these face representations. This has been proofed to be correct since the quality of a facial image was successfully predicted based on face embeddings [4], [18], [49].

In [38], Parde et al. analysed how well information about social-traits is retained in face representations. Human-assigned social trait profiles have been predicted using linear classifiers, in their experiments. They demonstrated that 11 social traits such as talkative, assertive, shy, quiet, warm, artistic, efficient, careless, impulsive, anxious, and lazy can be inferred from face embeddings to a high degree. The best-predicted traits included impulsive, warm, and anxious.

Hill et al. [19] analysed the representations of caricature faces. Their investigation included the categorization of viewpoint (0, 20, 30, 45, 60), illumination (ambient vs spotlight), gender (male vs female), and identity in embedding space. Their results conclude that information about face identity and imaging characteristics coexist, in a highly organized and hierarchical structure that is created by the utilized face recognition model. A summary of their results and a review about known properties of the face space, in the context of previous-generation face recognition algorithms, is given by O'Toole et al. [35].

In [56], [57] Zhong et al. conducted facial attribute estimation experiments using various mid-level representations from face recognition networks. By using various mid-level representations, they achieved highly accurate facial attribute estimation results. This indicates that also high-level representations, such as face recognition templates, might contain a significant amount of facial attribute information.

In [6], [9], [36], [47], the possibility of deriving demographic attributes such as gender, age, and race from face templates is demonstrated.

Previous works demonstrated that information about demographic attributes (e.g. gender, age, race), and social traits (e.g. impulsive, warm, and anxious), as well as head pose and image characteristics (e.g. quality, source of the image, viewpoint, illumination), can be derived from face templates. These works focused on the analysis of some specific attributes. The work of Terhörst et al. [43] provided a broader investigation on the predictability of over 100 attributes in face templates. Based on the prediction performance at two different reliability-levels, they categorized each attribute into one of three predictability classes. Their results demonstrate that up to 74 attributes can be accurately predicted from face templates.

In this work, we extend the analysis of [43] by providing a more in-depth investigation of attribute predictabilities. While in [43], the analysis is based on two difficulty-levels, we additionally provide experiments on a continuous difficulty range and extend the experiments to three different embedding types. This allows more fine-grained predictability statements of each attribute. We extend the analysis and discussion on the higher-level of attribute categories to derive more exact but also more compact and understandable findings. Moreover, we specifically discuss the implications of our results on future works.

3 Investigation methodology

The goal of this work is to analyse what attributes are stored in biometric face embeddings. We conduct this analysis by jointly training a classifier to accurately predict these attributes. If the classifier can accurately predict an attribute given the face embeddings, we can conclude that this attribute is encoded within the embedding. However, this investigation methodology only allows determining what attributes are stored in embeddings. It does not allow us to conclude what attributes are not encoded, since a reverse conclusion is not necessarily logical. If an estimator is not able to learn the pattern of an attribute, it does not imply that the pattern does not exist. The estimator might just not be able to deal with the complexity of the attribute pattern, or the data variability and representation might be low.

The following three subsections explain the different steps of our investigation methodology.

1) We explain the training procedure of our classifier. Training the classifier in a multi-task fashion allows making use of shared embedding space leading to general performance increases.
2) We explain the used methodology that allows the trained classifier to accurately state its predictions confidence (reliability).
3) We make use of this concept of prediction reliability to introduce predictability classes. These allow to more easily analyse the observations.

3.1 Massive attribute classifier (MAC)

The core of our attribute predictability analysis of face embeddings is a classification model. If this model is able to correctly predict an attribute given face embeddings, we can conclude that this attribute is encoded in the embeddings.

We trained a neural network model on face embeddings to jointly predict multiple attributes that might be stored

within. We refer to this model as a massive attribute classifier (MAC) due to the large number of attributes that are simultaneously learned. We evaluated multiple random network structures with 1-3 initial layers and 1-3 branch layers that connect the last initial layer with the softmax layers of each attribute. For each layer, a size of 128, 256, and 512 was evaluated. During this evaluation, we observed variations of the predicted performance per attribute by only 1-2%. Consequently, we decided on the network structure of the highest simplicity.

The chosen MAC-architecture builds on two initial layers, the input layer of size n_{in} (referring to the size of the used face embedding) and a second dense layer of size 512. The architecture makes use of a shared layer to increase the effectiveness of correlated attributes, such as age, gender, and race [14], [17]. While the multi-task MAC approach is highly-suitable for correlated attributes, for uncorrelated attributes, training single classifiers to predict each of these separately might, in some cases, lead to stronger attribute prediction performances as shown in [29]. Starting from the second layer, each attribute a has an own branch consisting of two additional layers of size 512 and $n_{out}^{(a)}$, where $n_{out}^{(a)}$ refers to the number of classes per attribute. For each layer a ReLU activation was used. The only exceptions are the output-layers that use softmax activations. Moreover, Batch-Normalization [21] and dropout [41] are applied to every layer. Using a dropout-strategy enables a more generalized performance and, more importantly, allows us to derive reliability statements about the prediction's confidence (described in Section 3.2). The quality of a reliability statement is robust to different magnitudes of dropout. Therefore, we followed the default dropout-probability of $p_{drop} = 0.5$ [41]. The training MAC-training was done in a multi-task learning fashion by applying a categorical cross-entropy loss for each attribute branch and use an equal weighting between each of these attribute-related losses. The training itself was based on an Adam optimizer [25] over $e = 200$ epochs with an initial learning rate $\alpha = 10^{-3}$ and a learning-rate decay of $\beta = \alpha/e$. The choices for these parameters as guided by the experiment setup of [47]. According to the amount of data available for training, the batch size b was chosen as $b = 1024$ for CelebA and $b = 16$ for LFW.

3.2 Prediction reliability

To formulate accurate predictions about the attribute-predictability in face embeddings, we make use of prediction reliabilities to simulate classifier circumstances of various difficulties. Following the methodology in [46], [47], we train the MAC with dropout. This allows us to state the MAC's prediction confidence (reliability). We perform $m = 100$ stochastic forward passes, to derive a reliability statement additionally to an attribute prediction. In each forward pass, a different dropout-pattern is applied, resulting in m different softmax outputs $v_i^{(a)}$ for each attribute a.

Given the outputs of the m stochastic forward passes of the predicted class $\hat{c}$ denoted as $x^{(a)} = v_{i,\hat{c}}^{(a)}$, the reliability measure is given as

$$rel(x^{(a)}) = \frac{1-\alpha}{m}\sum_{i=1}^{m} x_i^{(a)} - \frac{\alpha}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|x_i^{(a)} - x_j^{(a)}|,$$

with $\alpha = 0.5$, following the recommendation in [47]. The first part of the equation is a measure of centrality and utilizes the probability interpretation of the softmax output. A higher value can be interpreted as a high probability that the prediction is correct. The second part of the equation is the measure of dispersion and quantifies the agreement of the stochastic outputs x. This was shown to be an accurate reliability measure in [47].

We use this reliability measure to simulate the performance in circumstances of various difficulty. For each attribute, we calculate the prediction and corresponding reliability of each sample. Then, we sort the prediction according to their reliability, define a ratio of considered predictions (RCP), and compute the recognition performance based on this RCP-level. For instance, at 70% RCP the recognition performance is calculated on the predictions with the 70% of the highest reliabilities. Consequently, the performance at 100% RCP refers to the general performance of the whole dataset. An RCP-level of 100% refers to the most realistic, and thus challenging, circumstances. Lower RCP-levels will reject more predictions of low confidence that might contain factors of variances (such as blur and non-frontal head poses) that lead to unstable, and thus inaccurate, attribute estimates. Therefore, a low RCP-level refers to the MAC prediction performance under more optimal classifier circumstances.

Please note that also other predictability measures can be used for the proposed investigations. In [1], Alain et al. used linear separability to measure the predictability of a categorical attribute. However, if a binary attribute is perfectly encoded in the face space, the amount of information about this attribute does not change if the decision boundary in the embedding space is linear or curved. In [11], Dahr et al. measured the predictability based on the estimation of mutual information. While this approach does not rely on linear separability, it requires the training of additional networks to estimate the predictability. For these reasons, we choose to measure the predictability based on accurate prediction reliabilities [47] for our investigations.

3.3 Predictability classes

To derive more understandable statements about which attribute information is stored in a face embedding, we categorize each attribute into one of three predictability classes. These are based on the prediction performance at a RCP-level of 50% and 100%.

- **Easily-predictable (++):** an attribute is categorized as easily-predictable if, and only if, the balanced accuracy at 100% RCP is above 90%. This means that highly *accurate predictions are possible even under non-ideal circumstances* such as bad illuminations and non-frontal head poses.
- **Predictable (+):** an attribute is categorized as predictable if, and only if, the balanced accuracy at 100% RCP is under 90%, but the balanced accuracy at 50% RCP is above 90%. This indicates that highly *accurate predictions are possible under close-to-optimal conditions*, since it only takes into account 50% of the most confident MAC predictions.
- **Hardly-predictable (0):** an attribute is categorized as hardly-predictable if the balanced accuracy is below

90% at both, 100% and 50% RCP. Even *under close-to-optimal circumstances, the MAC is not able to reach high accuracies*. Consequently, the attribute patterns might be too complex for the MAC to handle or it does not exist a meaningful pattern for this attribute.

The attribute categories easily-predictable and predictable allow making confident statements about the amount of attribute information stored in face embeddings. However, this does not apply for hardly-predictable. If an attribute is categorized as hardly-predictable, the MAC is not able to accurately learn the pattern. This might have several reasons. First, the pattern does not exist. Second, the pattern does exist, but it is too complex for the model to learn. Or third, the pattern does exist but the amount of data and its representation is not appropriate for the classifier to learn. Consequently, for attributes categorized as hardly-predictable, we can not determine if a corresponding attribute pattern exists.

4 Experimental Setup

4.1 Databases

The Labeled Faces in the Wild (LFW) [20] and the CelebFaces Attributes (CelebA) [29] datasets provide a large number of attribute annotations and thus, are well suited to perform our predictability-analysis of the face space. Using a variety of soft-biometric labels, an in-depth investigation of which of these attributes are encoded in face embeddings is performed. Figure 1 shows sample images from both datasets. The CelebA dataset [29] covers more than 200k images taken from over 10k distinct celebrities. Each image is annotated with 40 binary attributes. Additionally, large variations in pose and background are covered. The LFW [20] dataset is comprised of 13k images taken from over 5k distinct individuals. Annotations for 73 binary attributes are provided for each image. Additionally, the images exhibit large variability in pose, lighting, focus, resolution, facial expression, age, gender, race, accessories, make-up, occlusions, background, and photographic quality. A wide range of characteristics (e.g. a person's demographics, skin, hair, beard, face geometry, periocular area, mouth, nose, accessories, and environment) are covered by the attribute labels of both databases [20], [29].

4.2 Cleaning attribute annotations of LFW

The attribute annotations of CelebA are of binary nature [29]. In contrast to CelebA, the attribute annotations contained in the LFW dataset are continuous and measure the degree of the attribute present in the image [20], [26], [27]. For instance, a high positive label score for the attribute beard should indicate a remarkable beard, while a negative annotation score indicates that no beard is shown. Consequently, binary labels can be derived by assigning true labels attributes with positive scores and false labels to attributes with negative label scores. However, a value around zero indicates that the attribute being present cannot be well determined.

To ensure a good performance of the MAC when trained on LFW, we manually converted the continuous attribute labels to binary labels. Using an upper and lower score threshold for each attribute, we assigned true labels to images with a score above the upper threshold, and false labels to images with a score below the lower threshold. Attributes with scores between the upper and lower score threshold bounds are labelled as undefined. The upper and lower thresholds for a particular attribute are manually determined by moving potential thresholds away from zero. At each candidate threshold, ten images with the closest attribute scores are investigated. By doing so, the original LFW annotations of the images are manually investigated for correctness. In the case that only eight or fewer images indicate that a particular attribute is present, the potential threshold is further moved away from the starting point until an adequate score threshold is found. If a potential threshold results in 9 or more correctly labelled images the threshold is chosen for that particular attribute. By repeating this procedure, the lower and upper thresholds for each of the attributes are identified. The scores are then binarized using the upper and lower thresholds to ensure an error-minimizing data basis of the MAC. This way, training and testing can be performed on meaningful and correctly labelled data.

Due to the generally poor quality of the LFW labels, our label-cleaning process reduces the number of used labels by 51,7%. This might induce a bias in our evaluation. To avoid biased conclusions that might result from this process, we evaluate another binary labelled database. After our label-cleaning process, we found 15 attribute labels of either a low number of positively and negatively labelled samples (<100). Table 1 provides a list of these attributes including the number of annotation divided into training and testing data. These attributes might have a low expressiveness in our facial analysis and thus, we will mark these attributes (in grey) in the following investigations.

TABLE 1: Train/test sample distribution on LFW for *selected attributes* that are found insufficient for a meaningful attribute analysis *after label-cleaning*. Pos and Neg refers to the number of positively and negatively labelled samples for the train and test set. The listed 15 attributes are found to be insignificant for the analysis due to a low number of samples in either the positive or negative class.

	Train		Test	
Attribute	Pos	Neg	Pos	Neg
Color Photo	8806	29	3772	24
Mouth Slightly Open	674	109	315	57
Round Face	9	588	3	250
Goatee	20	3346	10	1557
Baby	23	9137	15	3913
Bangs	89	5238	44	2080
Bald	114	4413	47	1953
Big Lips	101	751	48	318
Sunglasses	74	8583	50	3631
Partially Visible F.	124	1501	55	601
Mouth Wide Open	107	6593	56	2925
Double Chin	154	172	57	136
Harsh Lighting	113	914	62	487
Outdoor	173	510	63	243
Teeth Not Visible	125	2209	66	1089

Fig. 1: Sample images from CelebA (top row) and LFW (bottom row)

4.3 Evaluation metrics

Our predictability analysis of the face space is based on the prediction performance of the MAC. For calculating the prediction performance of facial attributes, the accuracy metric is usually used. However, this metric is defined by the ratio of correct predictions to the total number of predictions [33] and thus, is strongly affected in the case of unbalanced label-distribution. Therefore, we report the predictive performance in terms of balanced accuracy in order to be robust to attribute-imbalances. The balanced accuracy refers to the standard accuracy with class-balanced sample weights [23].

The datasets are divided into train/test data in a 70%/30% subject-exclusive split[1]. We decided not to use a cross-database evaluation protocol since both databases have over 30 non-overlapping attributes. Training on one database and evaluating on the other would result in the loss of this valuable attribute information. The predictive performance of a facial attribute estimator is analysed under circumstances of various difficulty. As described in Section 3.2, this is achieved by evaluating the prediction performance at various RCP-levels. While high RCP-levels simulate more realistic scenarios, low RCP-levels focus more on the confident predictions and thus, simulate more idealistic circumstances. This is used to determine more fine-grained statements about the attribute-predictabilities.

4.4 Face template extraction

For the experiments, we use three widely-used face recognition models based on FaceNet [40], CosFace [53], and ArcFace [10] losses. In this work, we use pre-trained models denoted as FaceNet[2], CosFace[3], and ArcFace[4]. The FaceNet and ArcFace model consist of a ResNet100 model trained on the MS1M database [16]. CosFace consists of a ResNet50 model trained on CASIA-WebFace [54]. Before providing the images as input for the models, the facial images are preprocessed (i.e. aligned, scaled, and cropped). The preprocessing for FaceNet is described in [22], for CosFace is described in [53], and for ArcFace is described in [15]. The embeddings are extracted by passing the preprocessed facial images to the face recognition models. The dimension of the obtained embeddings is 128 for FaceNet, 1024 for CosFace, and 512 for ArcFace.

1. Please note that attributes affected by imbalanced data training will be associated with a poorer prediction performance due to the use of balanced accuracies. Consequently, the imbalanced data training might lead to underestimating the amount of information stored in face embeddings for some attributes.
2. https://github.com/davidsandberg/facenet
3. https://github.com/MuggleWang/CosFace_pytorch
4. https://github.com/deepinsight/insightface

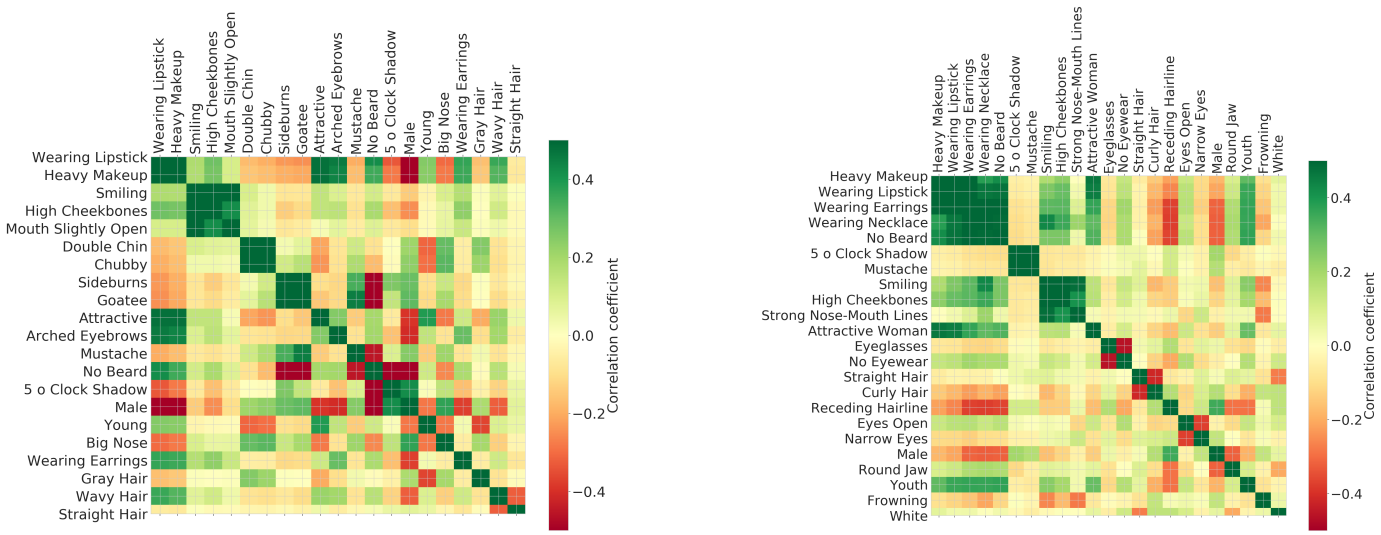
4.5 Investigations

This work targets to understand what information is stored in biometric face embeddings. To achieve this, we provide an in-depth investigation that is divided into the following parts:

1) We analyse the correlations between the attribute-annotations. The results of an attribute might show a high predictability which does not originate from the attribute information stored within an embedding but from correlated annotations of the testing database.
2) In two steps, we investigate which attributes are stored in face embeddings by analysing the attribute prediction performances. First, we analyse the prediction performance of each attribute on two specific confidence-levels of the MAC to get an overview of the problem. Second, we investigate the prediction performance of each attribute over a wide and continuous range of confidence-levels to achieve a more in-depth analysis of the stored information.
3) We compromise the detailed investigations to obtain an easily-understandable overview of which kind of information is encoded in face embeddings. First, we categorize each attribute into one of three predictability classes based on two-level prediction performances. Second, we visualize the predictability of each group of attributes to provide the reader with an intuitive understanding of which attributes are stored in face embeddings and how easily these can be predicted.

5 Results

Following the investigation plan from Section 4.5, this section works on the defined investigation points. Section 5.1 analysis the attribute correlations of the utilized face databases, Section 5.2 provides an in-depth investigation of the attribute predictability, and Section 5.3 summarizes the findings in a qualitatively and quantitatively manner. Finally, Section 5.4 discusses the implications for our findings on future works.



(a) CelebA (b) LFW

Fig. 2: Correlation of the attribute-annotations for CelebA and LFW. The attributes are chosen to show the 15 most positive and negative pairwise correlations. The attribute-correlation for LFW is shown after the label-cleaning process. Green indicate positive correlations, while red indicate a negative correlation. The correlation is based on the Pearson coefficient.

5.1 Attribute-correlation analysis

To avoid incorrect statements about which attributes are encoded in face embeddings, we first analyse the correlation of the attribute annotations. This aims at understanding the general label quality and avoids potential biases in the attribute annotations. Figure 2 shows a selection of attribute-label correlations for CelebA and LFW. The attributes are chosen to show the 15 most positive and negative pairwise correlations.

Figure 2a shows the attribute correlations for the annotations of CelebA. It can be seen that male faces do not correlate well with *Wearing Lipstick*, *Wearing Earrings*, and *Heavy Makeup*. These attributes correlate almost exclusively with female faces. Moreover, a large fraction of the male faces have a *Beard*, while this is not the case for women. Faces labelled as *Attractive* mostly belong to *Young* and *Female* faces wearing accessories and *Heavy Makeup*. Additionally, the figure approves the quality of some labels. For instance, the attribute *No Beard* negatively correlations with all types of beards such as *Goatee*, *Mustache*, and *Sideburns*.

The pairwise correlations of the attribute labels of LFW are shown in Figure 2b. The attributes *Heavy Makeup*, *Wearing Lipsticks*, *Wearing Earrings*, and *Wearing Necklace* belongs together with *Youth*, *Attractive Woman*, *Smiling*, and *High Cheekbones*. In contrast, these attributes do not correlate with *Receding Hairline* and *Male*. Similar as in Figure 2a, the correlation matrix in Figure 2b can be used to approve the label quality of some antagonistic attributes. For instance, *No Eyewear* negatively correlates with *Eyeglasses*, or *Curly Hair* negatively correlates with *Straight Hair*.

Since these attribute correlations can have an impact on the predictability investigation in Section 5.2, we additionally analysed the annotation correlation and the corresponding attribute prediction performance. The results analysing the 10 highest-correlating attribute pairs for CelebA and LFW are shown in Table 2. Given the attributes a and b, $\rho(a, b)$ represents the Pearson correlation coefficient. $a \mapsto b$ refers to the balanced accuracy when using the label of attribute a as the prediction for attribute b and vice versa. Generally, the highest correlations are found for the accessories *Wearing Lipstick*, *Wearing Earrings*, and *Heavy Makeup*. These attributes also show the highest prediction accuracies. If an attribute a is highly-predictable from face embeddings and there is a accurate correlation to attribute b ($a \mapsto b >$ 90%), then it can not be well differentiated if both attributes are encoded in the face embedding or just one of them. Consequently, these correlations have to be considered in the following evaluation.

5.2 Attribute-analysis of the face space

The investigation of which attributes are encoded in face embeddings is done based on the attribute prediction performance of the MAC. This is done in two degrees of detail. First, to get an overview, the prediction performance of the attributes is determined at two difficulty-levels. 100% RCP (hard) refers to the use of all samples under the given circumstances. 50% RCP (easy) refers to the 50% the predictions of which the classifier is most sure about its correctness. Second, the prediction performance of each attribute is analysed over a wide and continuous range of confidence-levels.

For CelebA, Table 3 shows the two-level prediction performance including the assigned predictability classes. Figures 3, 4 and 5 show the prediction performance at the continuous RCP range of $[0.5, 1]$ for all investigated face embeddings, FaceNet (FN), CosFace (CF), and ArcFace (AF). In general, two observations can be made. First, the prediction performance of lower RCP-levels is generally stronger than for higher RCP-levels. This demonstrates that the MAC learned to make reliable predictions of the CelebA dataset. Second, the prediction performance on FaceNet and CosFace is always slightly higher than the prediction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2021.3093920, IEEE Transactions on Biometrics, Behavior, and Identity Science

TABLE 2: Analysis of the annotation correlation and the corresponding attribute prediction performance: $a \mapsto b$ refers to the balanced accuracy when using the label of attribute a as the prediction for attribute b and vice versa. The correlation is given by the Pearson coefficient ρ. The 10 highest correlated attributes are investigated on both databases, CelebA and LFW. Only a few attribute correlations show strong influences on the prediction performance.

	Attribute a	Attribute b	$\rho(a,b)$	Accuracy $a \mapsto b$	Accuracy $b \mapsto a$
CelebA	Wearing Lipstick	Heavy Makeup	0.80	91.1%	89.1%
	Smiling	High Cheekbones	0.68	84.3%	84.0%
	Smiling	Mouth Open	0.54	76.8%	76.8%
	Double Chin	Chubby	0.53	74.2%	79.5%
	Sideburns	Goatee	0.51	74.4%	76.9%
	Wearing Lipstick	Attractive	0.48	74.0%	74.0%
	Wearing Lipstick	Arched Eyebrows	0.46	76.0%	70.4%
	Goatee	Mustache	0.45	77.4%	68.5%
	Wearing Lipstick	No Beard	0.42	78.2%	65.6%
	5 o Clock Shadow	Male	0.42	63.3%	82.8%
LFW	Heavy Makeup	Wearing Lipstick	0.64	77.7%	87.5%
	Wearing Lipstick	Wearing Earrings	0.60	71.1%	92.0%
	Wearing Earrings	Wearing Necklace	0.57	74.8%	83.2%
	5 o Clock Shadow	Mustache	0.55	85.0%	71.3%
	Smiling	High Cheekbones	0.54	86.5%	69.6%
	Heavy Makeup	Wearing Earrings	0.51	65.4%	91.4%
	Wearing Necklace	No Beard	0.49	75.4%	73.6%
	Strong No.-Mou. Lines	Smiling	0.48	76.8%	71.9%
	Heavy Makeup	Attractive Woman	0.46	79.8%	68.1%
	Wearing Lipstick	Attractive Woman	0.45	83.6%	65.1%

performance on ArcFace. The reason for this might be the large angular margin principle of ArcFace that distorts the feature space more incoherently and thus, makes it harder for estimators to learn existing patterns. In the other side, the embedding size seem to have less effect on the predictability since the smallest and largest embeddings (FaceNet-128, CosFace-1024) both achieve higher predictabilities than ArcFace (512). To summarize, many attributes from CelebA achieve a high prediction accuracy on all three face recognition models. This includes demographic characteristics, hairstyles, haircolors, and beardtypes. Additionally, the person's accessories are encoded with high details in the deeply-learned features.

For LFW, the two-level prediction performance including the assigned predictability classes is shown in Table 4. The grey highlights refer to results with limited validity since the label-cleaning process eliminated many samples with low-quality attribute annotations. The low number of train- and testing-samples might explain the weak prediction performance for some attributes, such as *Baby*, *Sunglasses*, and *Mouth*. Figures 3, 4 and 5 present the prediction performance at the continuous RCP range of $[0.5, 1]$ for all three embedding types. A lower RCP-level states a higher confidence of the classifier and thus, a higher balanced accuracy of the predicted attribute. For a few attributes, such as *Sideburns*, a counteracting behaviour is observed for low RCP-level. These might be explained by low annotation qualities of the ground truth [44]. Comparing the results of LFW with the results of CelebA shows that similar

TABLE 3: Prediction performance on CelebA: the performance is based on FaceNet (FN), CosFace (CN), and ArcFace (AF) embeddings and is reported in terms of balanced accuracies at two difficulty scenarios: 100% RCP (hard) and 50% RCP (easy). $^{++}$,$^{+}$, and ^{0} state the assigned predictability class.

		100% RCP			50% RCP		
	Attribute	FN	CF	AF	FN	CF	AF
Demo	Male^{++}	98.9%	97.1%	98.4%	99.9%	99.9%	99.9%
	Young^{+}	85.5%	82.7%	83.6%	96.4%	94.3%	94.5%
Skin	Pale Skin^{+}	76.0%	77.5%	71.9%	87.1%	90.0%	83.0%
	Rosy Cheeks^{+}	83.4%	85.8%	78.2%	96.3%	94.9%	81.7%
Hairstyle	Bald^{++}	95.7%	95.4%	94.0%	100.0%	100.0%	100.0%
	Bangs^{++}	91.7%	92.5%	89.3%	99.4%	99.6%	98.3%
	Receding Hairline^{+}	85.4%	84.6%	82.5%	96.4%	96.3%	94.2%
	Sideburns^{++}	92.8%	91.7%	92.1%	90.0%	96.2%	99.7%
	Straight Hair^{0}	68.6%	69.0%	70.7%	79.9%	80.0%	82.0%
	Wavy Hair^{0}	74.4%	74.5%	76.6%	86.4%	86.7%	89.4%
Haircolor	Black Hair^{+}	83.7%	84.5%	81.5%	96.6%	97.0%	94.3%
	Blond Hair^{++}	91.9%	91.8%	90.1%	99.3%	99.4%	98.3%
	Brown Hair^{+}	76.5%	78.2%	75.9%	90.1%	91.2%	88.3%
	Gray Hair^{++}	93.0%	92.9%	91.1%	99.6%	99.4%	98.8%
Beard	5 o Clock Shadow^{+}	86.9%	85.6%	85.8%	99.6%	99.2%	99.0%
	Goatee^{++}	93.4%	90.8%	91.8%	97.2%	100.0%	98.9%
	Moustache^{++}	92.2%	87.9%	89.7%	100.0%	94.4%	98.8%
	No Beard^{++}	92.1%	89.8%	90.8%	99.4%	99.4%	99.0%
Face Geo.	Chubby^{+}	86.5%	86.2%	83.1%	96.5%	97.4%	95.4%
	Double Chin^{+}	86.6%	87.4%	82.9%	96.9%	98.7%	95.4%
	High Cheekbones^{+}	78.5%	82.7%	72.2%	91.6%	95.0%	82.6%
	Oval Face^{0}	63.4%	64.6%	61.9%	70.8%	72.3%	68.1%
Periocular	Arched Eyebrows^{+}	79.8%	80.1%	77.0%	93.3%	93.6%	89.5%
	Bags Under Eyes^{0}	72.1%	74.6%	70.7%	80.6%	84.3%	80.7%
	Bushy Eyebrows^{+}	83.4%	83.1%	78.5%	95.9%	95.5%	91.9%
	Narrow Eyes^{0}	66.5%	70.2%	60.7%	75.4%	80.0%	66.7%
Mouth	Big Lips^{0}	74.6%	71.5%	68.8%	86.4%	83.7%	78.7%
	Mouth Slightly Open^{+}	74.5%	82.9%	67.5%	86.5%	95.5%	76.5%
	Smiling^{+}	80.1%	86.7%	71.7%	92.9%	97.7%	82.1%
Nose	Pointy Nose^{0}	71.7%	70.8%	69.3%	83.1%	83.1%	78.9%
	Big Nose^{0}	77.4%	76.7%	75.8%	88.1%	86.7%	87.1%
Accessories	Eyeglasses^{++}	97.3%	94.0%	90.6%	99.8%	99.7%	98.7%
	Heavy Makeup^{++}	90.1%	90.5%	88.7%	99.2%	99.5%	98.5%
	Wearing Earrings^{+}	79.2%	78.8%	77.0%	94.8%	93.6%	91.6%
	Wearing Hat^{++}	95.4%	95.1%	92.8%	99.4%	99.3%	99.0%
	Wearing Lipstick^{++}	92.8%	92.7%	91.4%	99.4%	99.7%	98.7%
	Wearing Necklace^{0}	71.8%	71.9%	71.4%	86.9%	86.5%	84.2%
	Wearing Necktie^{+}	83.7%	82.9%	82.1%	98.5%	98.1%	98.0%
Other	Blurry^{0}	74.3%	76.7%	68.2%	85.2%	89.4%	78.4%
	Attractive^{+}	79.6%	79.6%	77.9%	92.4%	92.4%	89.6%

prediction performances are achieved on attributes occurring in both datasets. Therefore, our label-cleaning process removed low-quality attribute-labels but did not result in a significant bias of the data. Due to the entangled patterns encoded in the templates some attributes, such as *Bold*, *Bangs*, and *Goatee*, are easy to learn and thus, achieve high performances. For some attributes, such as *High Cheekbones*, and *Smiling*, the MAC prediction performance is lower than for the single classifier approach reported in [29]. This

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2021.3093920, IEEE Transactions on Biometrics, Behavior, and Identity Science

TABLE 4: Prediction performance on LFW: the performance is based on FaceNet (FN), CosFace (CF), and ArcFace (AF) embeddings and is reported in terms of balanced accuracies at two difficulty scenarios: 100% RCP (hard) and 50% RCP (easy). $^{++}$,$^{+}$, and ^{0} state the assigned predictability class. Grey highlighting refers to reduced expressiveness due to limited data after the label-cleaning process.

		100% RCP			50% RCP		
	Attribute	FN	CN	AF	FN	CN	AF
Demographics	Male^{++}	98.3%	96.9%	83.9%	99.5%	99.6%	94.2%
	Baby^{0}	55.1%	49.9%	49.9%	50.0%	50.0%	50.0%
	Child^{0}	68.8%	73.1%	57.5%	75.8%	85.5%	52.4%
	Youth^{+}	79.9%	81.8%	70.5%	93.1%	94.4%	79.8%
	Middle Aged^{+}	88.4%	88.6%	74.0%	95.2%	97.9%	82.9%
	Senior^{++}	99.6%	97.8%	83.9%	100.0%	100.0%	88.4%
	Asian^{++}	95.5%	90.4%	66.2%	100.0%	97.3%	69.6%
	White^{++}	97.4%	94.4%	73.6%	99.4%	99.1%	81.4%
	Black^{++}	95.3%	92.3%	63.2%	98.3%	100.0%	53.6%
	Indian^{+}	85.2%	63.0%	50.2%	92.5%	50.0%	54.7%
Skin	Rosy Cheeks^{0}	67.2%	71.0%	58.8%	73.0%	77.1%	64.3%
	Shiny Skin^{+}	82.1%	89.4%	67.9%	89.7%	99.9%	75.6%
	Pale Skin^{0}	68.0%	73.1%	62.9%	79.9%	83.2%	67.2%
	Flushed Face^{0}	66.5%	73.9%	55.5%	77.5%	77.5%	52.3%
Hairstyle	Curly Hair^{0}	69.0%	72.6%	61.7%	77.8%	83.5%	68.7%
	Wavy Hair^{++}	95.0%	96.7%	80.5%	99.7%	99.7%	83.3%
	Straight Hair	67.5%	69.5%	59.8%	76.8%	80.0%	65.5%
	Receding Hairline^{+}	83.3%	83.9%	73.0%	93.5%	95.0%	84.9%
	Bangs^{++}	97.0%	94.9%	64.1%	100.0%	100.0%	50.0%
	Bald^{++}	93.6%	84.2%	75.8%	97.9%	96.4%	75.0%
	Sideburns^{++}	98.9%	98.5%	84.1%	99.7%	99.7%	89.2%
Haircolor	Black Hair^{++}	90.4%	89.0%	65.6%	96.5%	96.4%	61.5%
	Blond Hair^{++}	95.2%	94.6%	71.7%	98.8%	100.0%	55.6%
	Brown Hair^{+}	81.5%	84.1%	71.9%	91.9%	95.3%	82.7%
	Gray Hair^{++}	98.8%	96.5%	88.4%	100.0%	100.0%	93.9%
Beard	No Beard^{++}	98.1%	94.9%	83.9%	100.0%	100.0%	92.1%
	Moustache^{++}	98.5%	93.7%	79.7%	99.3%	96.8%	78.1%
	5 o Clock Shadow^{++}	96.5%	95.7%	83.8%	99.6%	99.7%	92.4%
	Goatee^{++}	94.5%	84.8%	70.0%	100.0%	100.0%	100.0%
Face Geometry	Oval Face^{++}	82.7%	90.4%	71.6%	95.4%	96.8%	75.8%
	Square Face^{++}	99.1%	96.3%	89.1%	100.0%	99.6%	96.3%
	Round Face^{+}	84.2%	71.4%	49.6%	100.0%	100.0%	50.0%
	Round Jaw^{+}	70.6%	84.7%	60.8%	81.1%	95.0%	58.4%
	Double Chin^{++}	91.5%	96.0%	81.1%	100.0%	100.0%	88.7%
	High Cheekbones^{++}	79.9%	96.9%	73.3%	90.4%	99.9%	81.8%
	Chubby^{+}	85.5%	86.1%	74.3%	98.0%	97.5%	79.4%
	Obstructed Forehead^{++}	85.9%	93.2%	65.0%	99.9%	98.3%	61.3%
	Partially Visible F.$^{+}$	85.2%	85.0%	65.9%	94.0%	95.7%	50.0%
	Fully Visible F.$^{+}$	85.9%	88.7%	71.8%	95.4%	98.2%	82.2%

		100% RCP			50% RCP		
	Attribute	FN	CN	AF	FN	CN	AF
Periocular	Eyes Open^{0}	60.4%	70.9%	54.4%	63.6%	71.7%	54.8%
	Brown Eyes^{+}	82.1%	84.8%	64.0%	92.8%	93.9%	66.8%
	Bags Under Eyes^{++}	87.2%	93.3%	73.7%	95.4%	98.3%	83.5%
	Narrow Eyes^{+}	77.1%	83.2%	66.2%	86.3%	92.3%	74.1%
	Bushy Eyebrows^{++}	96.3%	95.7%	83.8%	99.1%	98.8%	91.7%
	Arched Eyebrows^{+}	85.3%	86.6%	71.6%	94.5%	96.1%	76.8%
Mouth	Mouth Closed^{+}	73.2%	85.0%	64.0%	83.9%	95.9%	72.4%
	Mouth Slightly Open^{+}	73.8%	89.1%	61.8%	83.0%	96.6%	65.1%
	Mouth Wide Open^{+}	66.6%	85.5%	50.8%	59.9%	90.9%	50.0%
	Teeth Not Visible^{+}	70.0%	84.8%	65.2%	75.3%	99.8%	58.3%
	Smiling^{++}	72.0%	93.8%	67.9%	81.3%	99.7%	75.9%
	Big Lips^{++}	87.6%	92.5%	57.3%	98.0%	92.3%	57.8%
Nose	Big Nose^{+}	84.5%	88.8%	71.6%	93.6%	97.3%	81.5%
	Pointy Nose^{++}	96.5%	95.4%	71.5%	100.0%	100.0%	71.3%
	Str. No.-Mou. Lines^{++}	70.0%	94.2%	61.7%	80.7%	99.3%	71.6%
Accessories	Heavy Makeup^{++}	96.7%	96.3%	69.9%	99.0%	100.0%	57.1%
	Wearing Hat^{++}	87.2%	91.7%	67.9%	96.9%	98.3%	53.8%
	Wearing Earrings^{++}	91.7%	91.0%	73.3%	97.9%	97.8%	72.9%
	Wearing Necktie^{+}	84.6%	81.5%	72.8%	93.5%	91.1%	75.2%
	Wearing Necklace^{+}	83.7%	86.0%	74.1%	92.1%	95.1%	82.5%
	Wearing Lipstick^{++}	98.5%	99.1%	75.9%	99.5%	100.0%	74.0%
	No Eyewear^{++}	95.5%	90.4%	86.1%	98.2%	97.7%	90.3%
	Eyeglasses^{++}	96.1%	87.6%	90.0%	98.4%	97.3%	95.6%
	Sunglasses^{+}	71.6%	82.7%	50.8%	62.4%	100.0%	50.0%
Environment	Blurry	61.4%	78.6%	57.2%	66.3%	89.5%	58.6%
	Harsh Lighting^{+}	76.0%	87.3%	61.3%	89.1%	90.8%	57.9%
	Flash^{++}	78.3%	92.6%	58.3%	88.3%	98.8%	51.5%
	Soft Lighting	65.7%	73.8%	60.2%	72.3%	84.8%	66.1%
	Outdoor^{+}	77.2%	88.8%	60.8%	81.9%	97.0%	65.9%
Other	Frowning^{++}	78.3%	97.4%	72.4%	88.8%	99.9%	79.5%
	Color Photo^{0}	72.8%	70.6%	54.0%	75.0%	50.0%	60.0%
	Posed Photo^{+}	76.0%	88.3%	60.7%	80.9%	98.4%	63.0%
	Attractive Man^{0}	74.4%	75.0%	65.0%	85.1%	85.9%	74.2%
	Attractive Woman^{++}	95.3%	95.7%	75.1%	100.0%	98.6%	71.4%

shows that a single classifier with higher capacity trained on these attributes might result in stronger prediction performances than the MAC approach. In general, the prediction performance on FaceNet and CosFace is stronger than on ArcFace. Due to the large angular margin principle, ArcFace embeddings contain more complex attribute patterns. For the experiments on LFW, less data was available for training, since we needed to filter low-quality labels to guarantee a high validity of the results. Therefore, it can be expected that the performance on ArcFace might be higher if with more training data is available. Similarly to FaceNet, many soft-biometric attributes are strongly encoded in CosFace embeddings. Moreover, attributes belonging to the categories *Mouth* and *Environment* show a much higher predictability for CosFace than for FaceNet. In contrast, only some attribute categories, such as *Haircolor*, *Hairstyle*, *Accessories*, and *Beard*, show a high predictability on ArcFace embeddings. CosFace and ArcFace are both margin-based losses. However, only the additive angular margin loss ArcFace showed reduced predictability results in comparison to triplet-loss and CosFace. This demonstrates an effect of the training loss on the attribute predictability. Moreover, the effect of the training loss might be stronger than a potential effect of the embedding size on the attribute predictability since the lowest- and highest-dimensional embeddings (FaceNet 128, CosFace 1024) both performed well in pre-

dicting soft-biometric attributes while the prediction performance on ArcFace embeddings of 512 dimensions is smaller. Nevertheless, many attributes can be predicted from the embeddings with a high degree of reliability. This can be observed for demographics, hairstyles, haircolors, beard types, accessories. However, as demonstrated in Section 5.1, there is a high correlation between the accessories *Wearing Lipstick* and *Heavy Makeup* that strongly affects the prediction performance of the MAC. Consequently, for these attributes, it can not be well differentiated if both are encoded in the face embedding or just one of these. Additionally, characteristics about the face geometry such as face shape, the presences of a double chin, and forehead visibility can be determined. Attributes that do not directly belong to the user, such as lighting conditions or image blurriness, could not reliably predicted with the MAC. It should be noted that the high correlation to various accessories might lead to the high predictability of *Attractive Woman*.

5.3 Category-wise analysis of the face space

In the previous section, we discussed the results on the level of single attributes. We showed that from the 113 investigated attributes, 39 attributes belong to the class easily-predictable, 35 belong to class predictable and 39 belong to the class hardly-predictable. In this section, we discuss the findings in a coarse to fine manner and on a more abstract level, the level of attribute categories.

Table 5 summarizes the categories of the attributes in these three predictability classes, to obtain a more general overview of the encoded information in the face embeddings. To provide a more complete view of the problem, this table also includes observations from related works, such as findings about head pose [37] and image quality [4]. Although, face recognition models are trained for recognition, attributes that belong to the categories such as *Face Geometry*, *Periocular Area*, *Nose*, and *Mouth*, are not easily-predictable. In contrast, non-permanent factors that modern face recognition systems aim to be robust at, turn out to be easily-predictable. For instance, this includes *Hairstyles*, *Haircolors*, *Beards*, *Accessories*, *Head Poses*, and *Social Traits*

TABLE 5: Categorized summary of the predictability classes including findings of related works.

Easily-predictable	Predictable	Hardly-predictable
Demographics	Face Geometry	Skin
Hairstyle	Periocular	Mouth
Haircolor	Nose	Environment
Beard	Image Quality [4]	
Accessories		
Head Pose [37]		
Social Traits [38]		

Figure 6 provides a more detailed predictability-overview of the attribute categories. The analysis is divided between both investigated face embeddings. On the axes, the prediction performance of two RCP-levels are shown. Each figure is divided into three areas representing the three predictability classes. The grey area represents the hardly-predictable class (0), the light green area represents the predictable class (+), and the dark green area represents the easily-predictable class (++). Moreover, each point indicate the average performance of the attributes belonging to the attribute-category. The elliptic shaded area around each point indicates the (standard) deviation of individual performance of the corresponding attributes. The x-axis of the shaded area represents the standard deviation of the performance at 100% RCP (more realistic circumstances), while the y-axis of the shaded are represents the deviation of the performance at 50% RCP (more idealistic circumstances).

In Figures 6a and 6b, the predictability of the attribute-categories are shown for FaceNet. Figures 6c and 6d show the predictability of the attribute-categories for CosFace and for ArcFace, the predictability of the attribute-categories are shown in Figures 6e and 6f. It can be seen that many attribute-categories are richly encoded in the FaceNet and CosFace embeddings. This includes different *Haircolors*, *Hairstyles*, *Beards*, *Accessories*, and *Demographics*, as well as attributes that belong to the *Face Geometry*, the *Nose* and *Periocular* area. For ArcFace, it can be seen that more attribute-categories belong to the grey (hardly-predictable) area. Due to the large angular margin principle of ArcFace, the face embeddings contain attribute patterns of higher complexity. The reduced amount of training data combined with the additive angular margin loss of ArcFace might be one of the reasons that many attribute categories belong to the hardly-predictable class. Both, the reduced amount of training data due to the label cleaning process as well as the more complex attribute pattern due to the ArcFace loss might it more challenging for the MAC to accurate predict attributes. However, the large elliptic shades in the grey areas indicate that these categories possess some highly predictable attributes as well. The high-level view on the attribute-categories lead to some valuable information loss in order to simplify the relations. Our investigation methodology can only state what information is stored in biometric face embeddings, but does not allow statements about what attributes are not encoded. Consequently, we can only surely conclude about four attribute-categories. The attributes *Haircolor*, *Hairstyle*, *Beard*, and *Accessories* are strongly encoded in ArcFace embeddings.

The reason that face recognition networks tend to retain soft-biometric information might lie in their correlation to their user's identity. Recent works [2], [39], [44] showed that soft-biometric attributes of a face provide enough information to be successful applied in verification and identification tasks. Consequently, there is a strong link between these attributes, the appearance of a person and its identity. This link might be the reason that deep networks trained for identification retain these attributes.

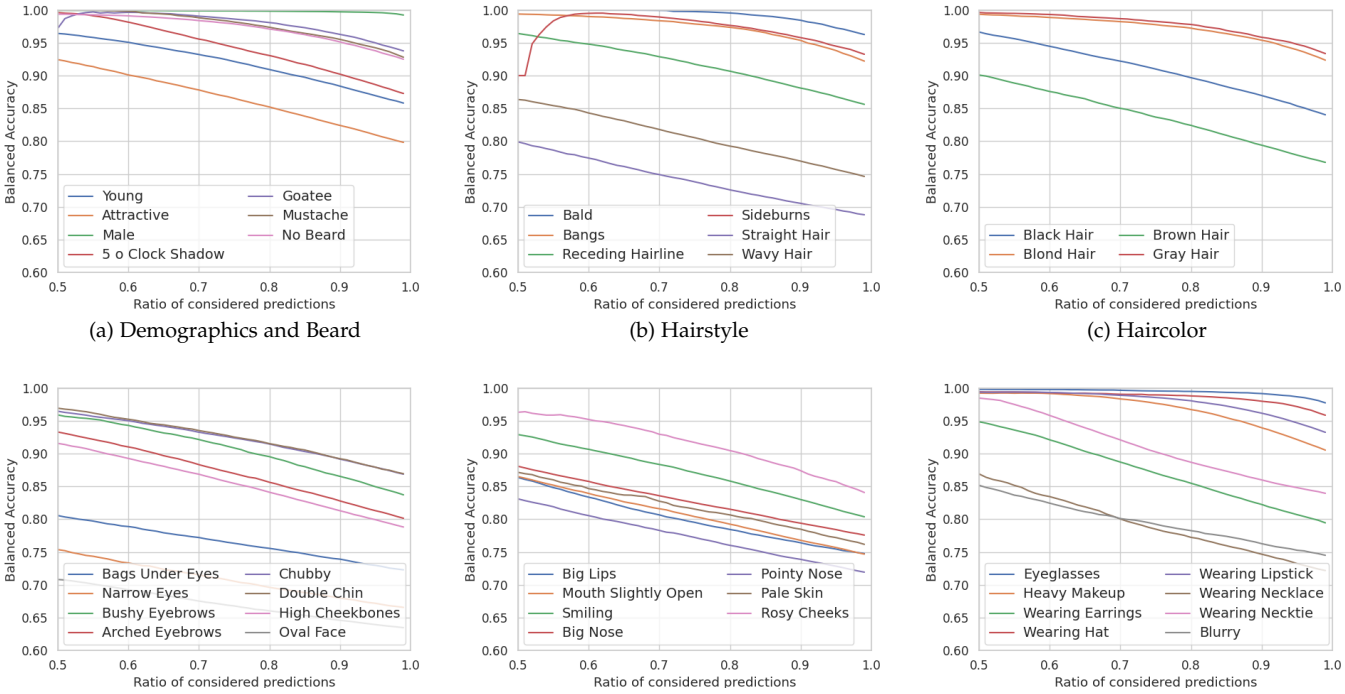
5.4 Implications of our findings

The findings of this work might have important consequences for future research in privacy-preserving and bias-mitigating face recognition.

5.4.1 Privacy in face recognition

The experiments demonstrated major privacy-risks in face recognition systems. For many applications, the user of a face recognition system provides his/her biometric data solely for recognition. The embeddings extracted from a face

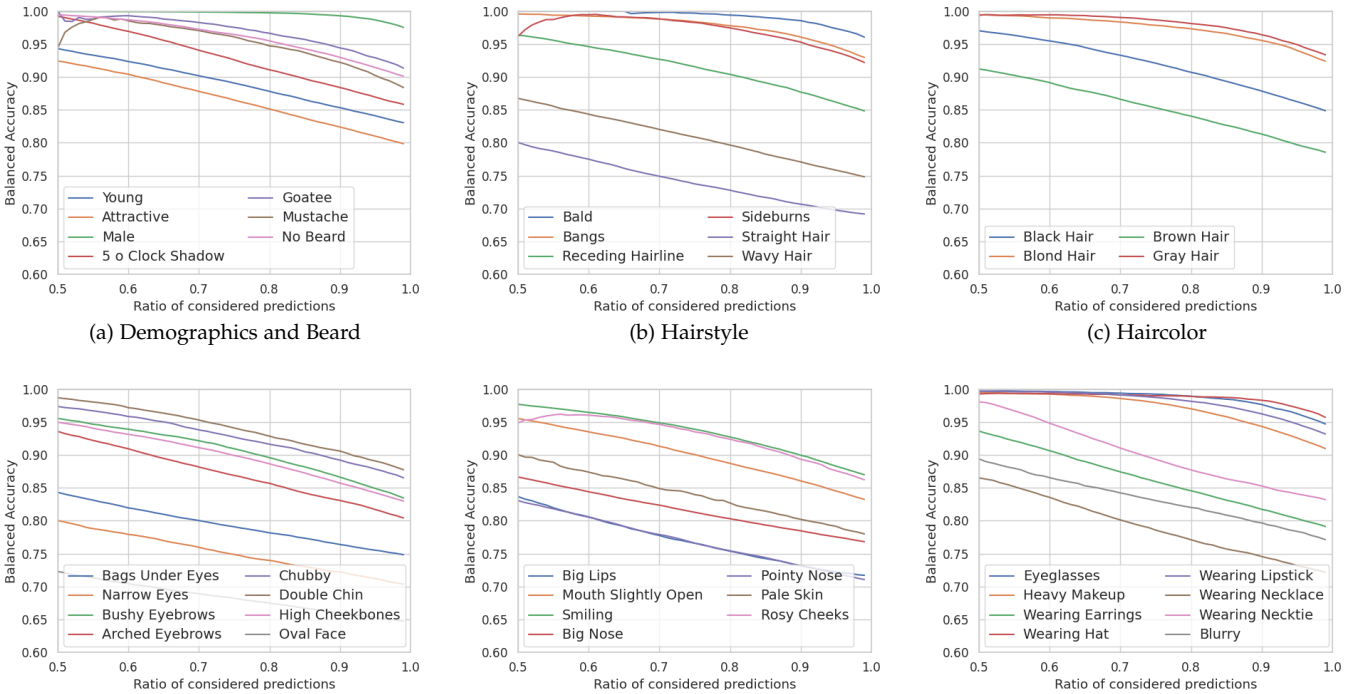
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2021.3093920, IEEE Transactions on Biometrics, Behavior, and Identity Science



(a) Demographics and Beard (b) Hairstyle (c) Haircolor

(d) Periocular and Face Geometry (e) Mouth, Nose, and Skin (f) Accessories and Environment

Fig. 3: Accuracy-Reliability plots for the CelebA database on FaceNet embeddings. The balanced accuracy of the MAC is shown for a continuous RCP range of [0.5, 1]. The MAC performance of the 40 attributes is divided into 6 categories represented by subfigures (a)-(f) to allow a simple category-based analysis.

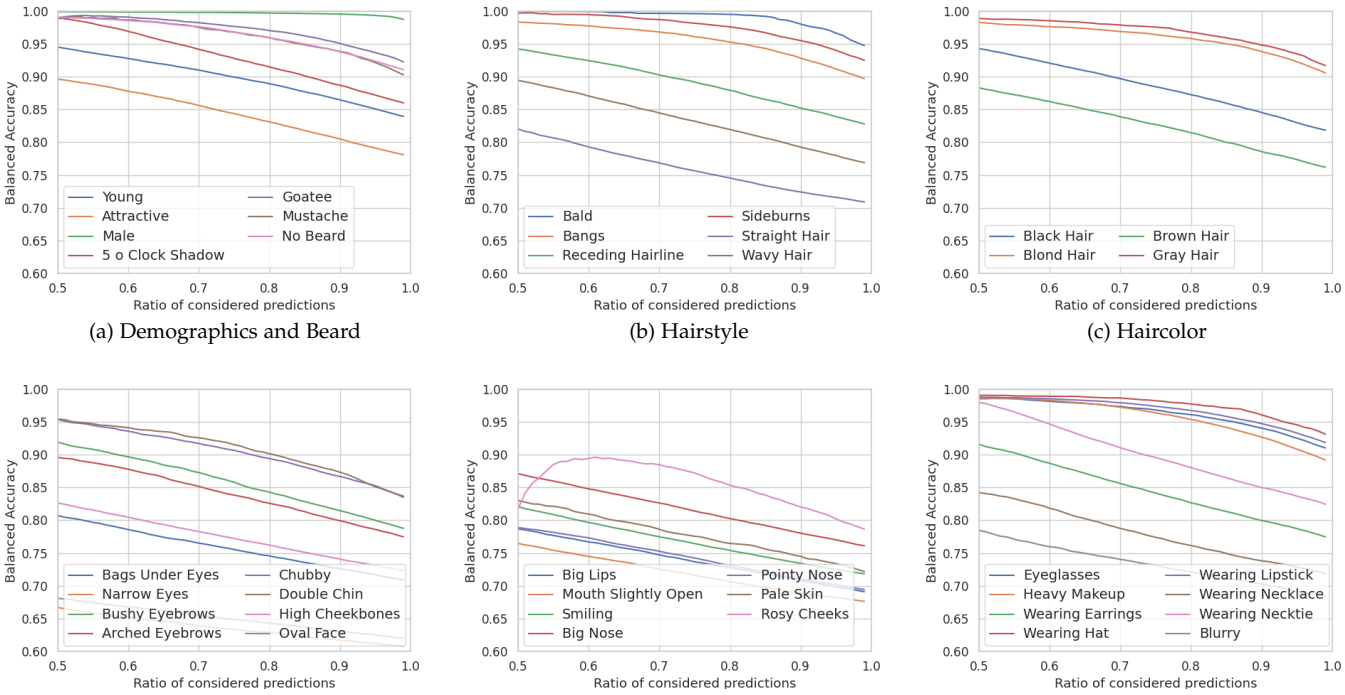


(a) Demographics and Beard (b) Hairstyle (c) Haircolor

(d) Periocular and Face Geometry (e) Mouth, Nose, and Skin (f) Accessories and Environment

Fig. 4: Accuracy-Reliability plots for the CelebA database on CosFace embeddings. The balanced accuracy of the MAC is shown for a continuous RCP range of [0.5, 1]. The MAC performance of the 40 attributes is divided into 6 categories represented by subfigures (a)-(f) to allow a simple category-based analysis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2021.3093920, IEEE Transactions on Biometrics, Behavior, and Identity Science



(a) Demographics and Beard (b) Hairstyle (c) Haircolor

(d) Periocular and Face Geometry (e) Mouth, Nose, and Skin (f) Accessories and Environment

Fig. 5: Accuracy-Reliability plots for the CelebA database on ArcFace embeddings. The balanced accuracy of the MAC is shown for a continuous RCP range of [0.5, 1]. The MAC performance of the 40 attributes is divided into 6 categories represented by subfigures (a)-(f) to allow a simple category-based analysis.

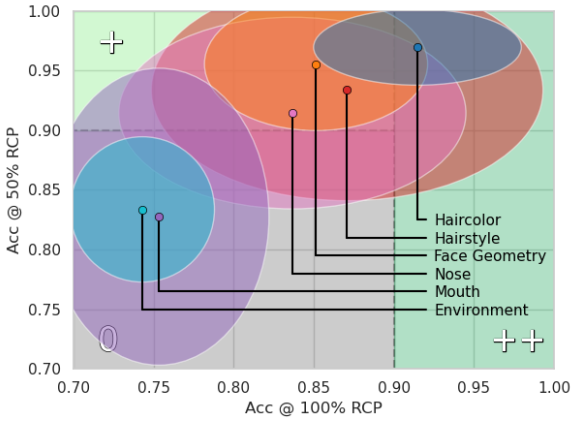
recognition systems should only contain identity-related information in order to prevent a potential misuse (function creep) of this private data. However, our experiments demonstrated that face embeddings also contain information of privacy-sensitive attributes, raising major privacy-risks. Consequently, future works have to deal with these privacy-issues, for instance by providing solutions to suppress attribute information in face embeddings.

5.4.2 *Bias in face recognition*

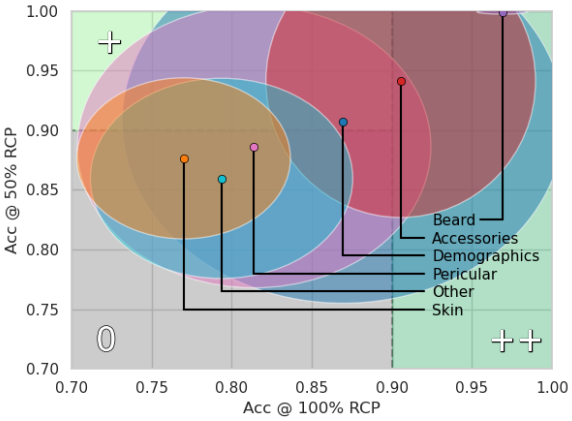
Many attributes are encoded in face embeddings as our experiments have shown. Although face recognition embeddings are trained to be robust against non-permanent factors, the results demonstrate that especially these attributes are accurately predictable from face templates. This includes information about *Hairstyles*, *Haircolors*, *Beards*, and *Accessories* for ArcFace and more attributes for FaceNet. The existence of these attribute-traits in face embeddings indicates that current face recognition systems are still not robust to these non-demographic factors as it was shown in recent works [3], [7], [50]. Consequently, future works have to propose solutions to also mitigate non-demographic bias in face recognition.

6 Conclusion

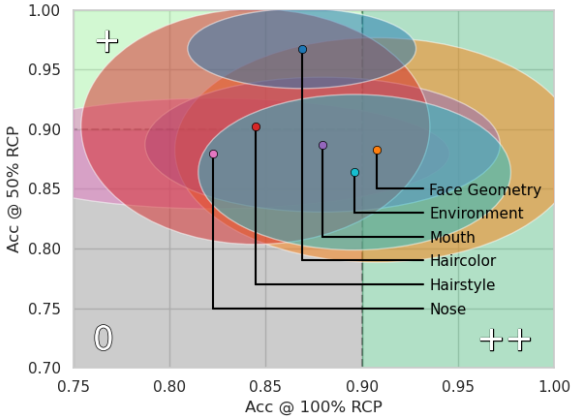
The current success of face recognition systems is driven by the advances of deeply-learned face embeddings. However, these embeddings contain more information then just the person's identity as recent works have shown. For instance, demographics, image characteristics, and social traits are additionally encoded in these embeddings. This might lead to biased decisions in face recognition systems and raises major privacy issues. To mitigate these privacy and bias concerns, deep knowledge about the encoded information in face embeddings is needed. Consequently, in this work, we provide a more in-depth analysis of what information is stored in biometric face embeddings. We analysed 73 different soft-biometric attributes towards their predictability from three popular face embeddings over a wide range of difficulty-levels. To enhance the understandability of the results, we additionally investigated the predictability of several categories of attributes. This was done assigning each group into one of three predictability classes as well as by analysing the predictability in a continuous range. The results demonstrate the many attributes are encoded in biometric face embeddings. About one-third of the analysed attributes can be classified as easily-predictable, another third as predictable, and one-third is only hardly-predictable from face embeddings. We could show that especially attributes related to haircolor, hairstyles, beards, and accessories are strongly encoded in face embeddings. Although that face recognition templates are trained to be robust against non-permanent factors, we demonstrated that specifically these attributes are easily-predictable from face embeddings. We hope that future works build on the knowledge of this work to develop accurate face recognition solutions that additionally focuses on mitigating bias and privacy concerns of various origins.



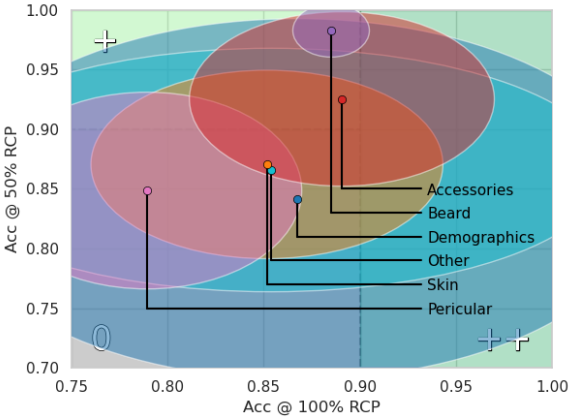
(a) FaceNet - Part 1/2



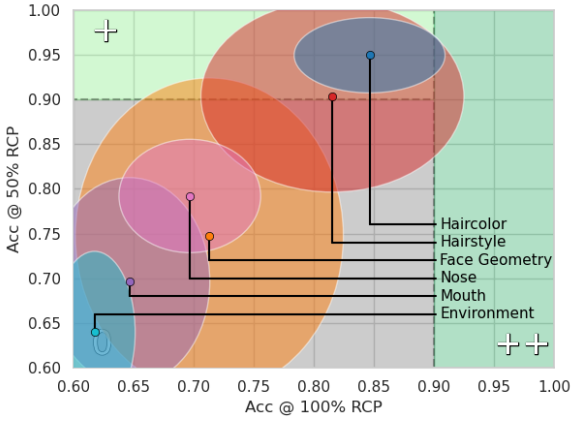
(b) FaceNet - Part 2/2



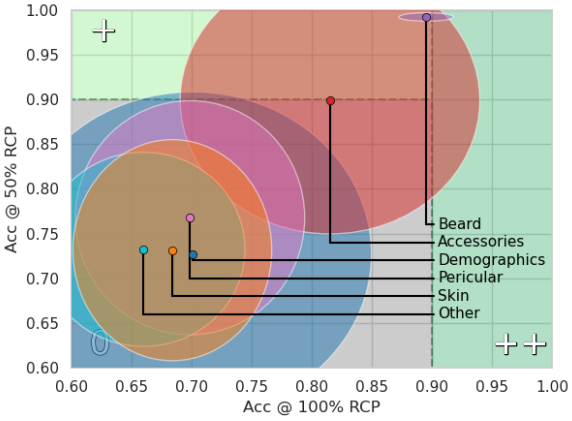
(c) CosFace - Part 1/2



(d) CosFace - Part 2/2



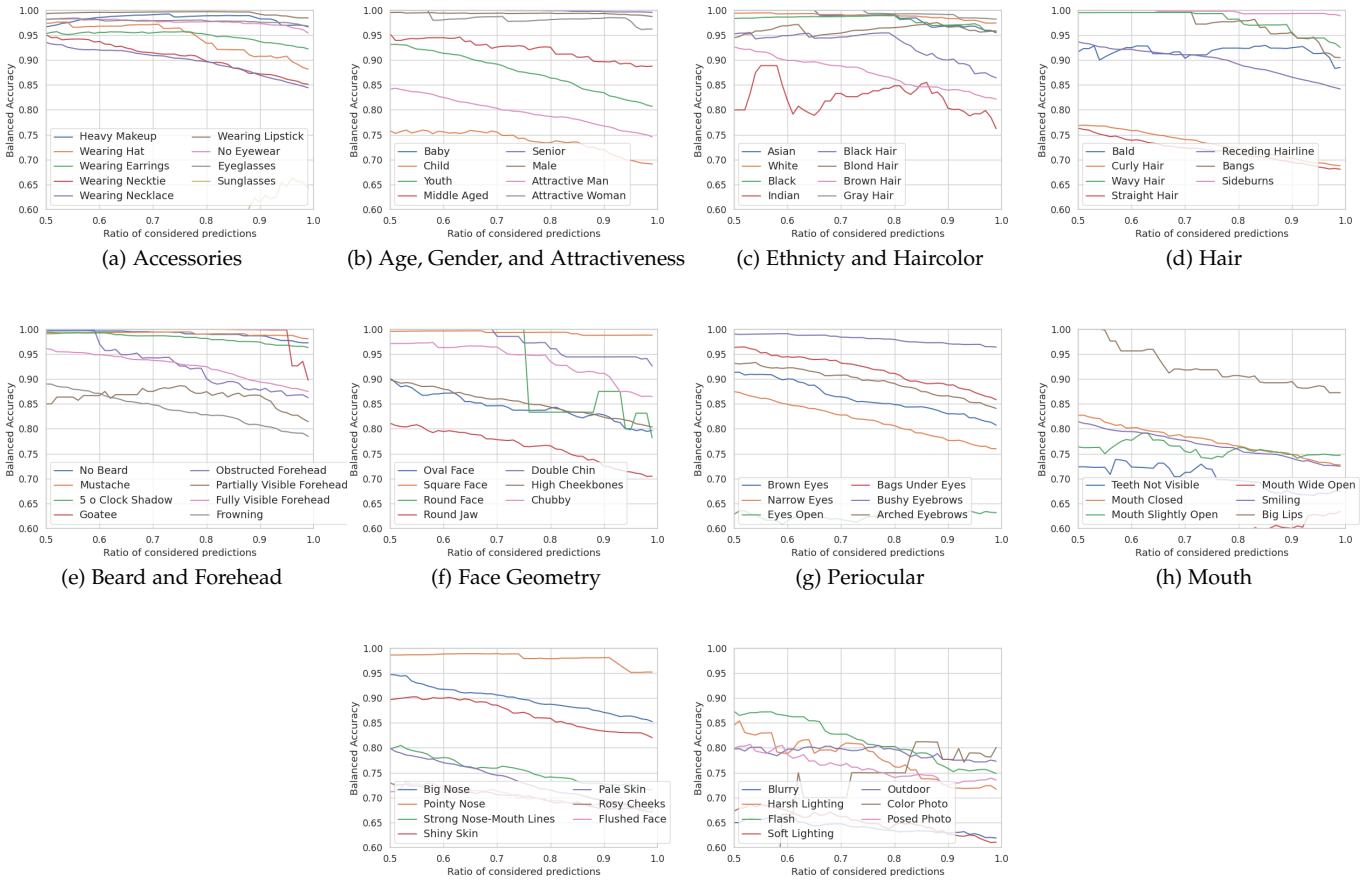
(e) ArcFace - Part 1/2



(f) ArcFace - Part 2/2

Fig. 6: Visual summary of the categorized attribute predictability. The axes represent the balanced prediction accuracy at two RCP-levels. The figures are divided into three areas representing the three predictability-classes. The dark green are indicates the class easily-predictable (++), the light green area indicates the class predictable (+), and the grey area indicates the hardly-predictable (0). Each point represent the average performance of an attribute category. The shaded area around each point represents the (standard) deviation of the individual attribute-performances belonging to the category. Many attributes are highly-encoded in the face embeddings.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2021.3093920, IEEE Transactions on Biometrics, Behavior, and Identity Science



(a) Accessories (b) Age, Gender, and Attractiveness (c) Ethnicity and Haircolor (d) Hair

(e) Beard and Forehead (f) Face Geometry (g) Periocular (h) Mouth

(i) Nose and Skin (j) Others

Fig. 7: Accuracy-Reliability plots for the LFW database on FaceNet embeddings. The balanced accuracy of the MAC is shown for continuous RCP range of [0.5, 1]. The MAC performance of the 73 attributes is divided into 10 categories represented by subfigures (a)-(j) to allow a simple category-based analysis.

ACKNOWLEDGEMENTS

This work was supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within the National Research Center for Applied Cybersecurity (ATHENE), and in part by the German Federal Ministry of Education and Research (BMBF) through the Software Campus project.

REFERENCES

[1] G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[2] N. Almudhahka, M. S. Nixon, and J. S. Hare. Human face identification via comparative soft biometrics. In *IEEE International Conference on Identity, Security and Behavior Analysis, ISBA 2016, Sendai, Japan, February 29 - March 2, 2016*, pages 1–6. IEEE, 2016.

[3] G. Balakrishnan, Y. Xiong, W. Xia, and P. Perona. Towards causal benchmarking of bias in face analysis algorithms. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII*, volume 12363 of *Lecture Notes in Computer Science*, pages 547–563. Springer, 2020.

[4] L. Best-Rowden and A. K. Jain. Learning face image quality from human assessments. *IEEE Transactions on Information Forensics and Security*, 13(12):3064–3077, Dec 2018.

[5] B. Bortolato, M. Ivanovska, P. Rot, J. Krizaj, P. Terhörst, N. Damer, P. Peer, and V. Struc. Learning privacy-enhancing face representations through feature disentanglement. In *15th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2020, Buenos Aires, Argentina, May 18-22, 2020*. IEEE, 2020.

[6] F. Boutros, N. Damer, P. Terhörst, F. Kirchbuchner, and A. Kuijper. Exploring the channels of multiple color spaces for age and gender estimation from face images. In *22th International Conference on Information Fusion, FUSION 2019, Ottawa, ON, Canada, July 2-5, 2019*, pages 1–8. IEEE, 2019.

[7] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Trans. Biom. Behav. Identity Sci.*, 1(1):32–41, 2019.

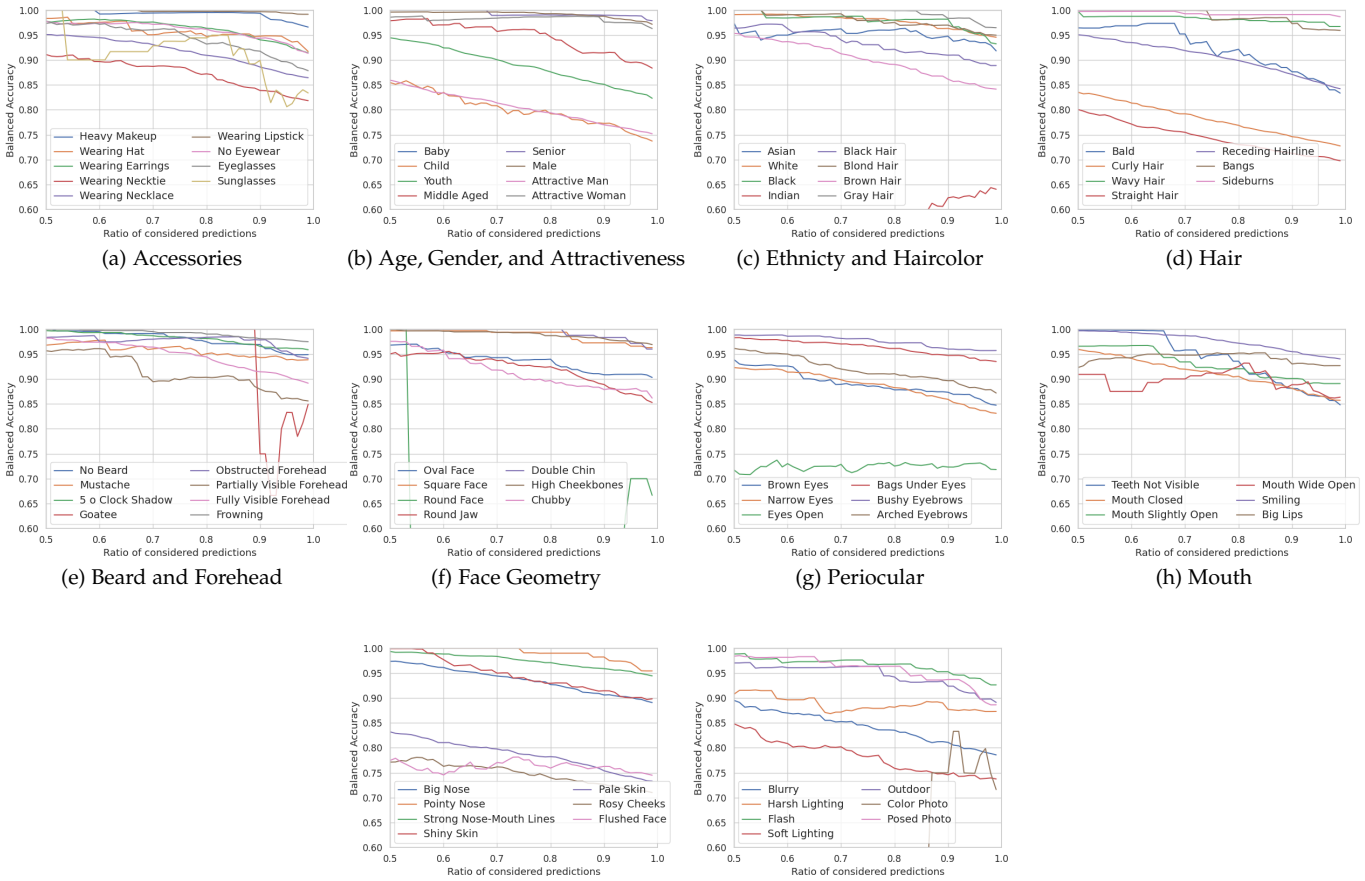
[8] N. Damer, Y. Wainakh, V. Boller, S. von den Berken, P. Terhörst, A. Braun, and A. Kuijper. Crazyfaces: Unassisted circumvention of watchlist face identification. In *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*, pages 1–9. IEEE, 2018.

[9] A. Das, A. Dantcheva, and F. Bremond. Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. In L. Leal-Taixé and S. Roth, editors, *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11129 of *Lecture Notes in Computer Science*, pages 573–585. Springer, 2018.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] P. Dhar, A. Bansal, C. D. Castillo, J. Gleason, P. J. Phillips, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2021.3093920, IEEE Transactions on Biometrics, Behavior, and Identity Science



(a) Accessories (b) Age, Gender, and Attractiveness (c) Ethnicity and Haircolor (d) Hair

(e) Beard and Forehead (f) Face Geometry (g) Periocular (h) Mouth

(i) Nose and Skin (j) Others

Fig. 8: Accuracy-Reliability plots for the LFW database on CosFace embeddings. The balanced accuracy of the MAC is shown for continuous RCP range of [0.5, 1]. The MAC performance of the 73 attributes is divided into 10 categories represented by subfigures (a)-(j) to allow a simple category-based analysis.

R. Chellappa. How are attributes expressed in face dcnns? In *15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020, Buenos Aires, Argentina, November 16-20, 2020*, pages 85–92. IEEE, 2020.

[12] S. Gong, X. Liu, and A. K. Jain. DebFace: De-biasing face recognition. *CoRR*, abs/1911.08080, 2019.

[13] P. Grother, M. Ngan, and K. Hanaoka. Ongoing face recognition vendor test (frvt) part 2: Identification. *NIST Interagency/Internal Report (NISTIR)*, 2018.

[14] G. Guo and G. Mu. Joint estimation of age, gender and ethnicity: CCA vs. PLS. In *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, Shanghai, China, 22-26 April, 2013*, pages 1–6. IEEE Computer Society, 2013.

[15] J. Guo, J. Deng, N. Xue, and S. Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 44. BMVA Press, 2018.

[16] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2016.

[17] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(11):2597–2609, 2018.

[18] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, R. Haraksim, and L. Beslay. FaceQnet: Quality assessment for face recognition based on deep learning. In *IEEE International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, Jun. 2019.

[19] M. Q. Hill, C. J. Parde, C. D. Castillo, Y. I. Colon, R. Ranjan, J. Chen, V. Blanz, and A. J. O'Toole. Deep convolutional neural networks in the face of caricature: Identity and image revealed. *CoRR*, abs/1812.10902, 2018.

[20] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[21] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org, 2015.

[22] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1867–1874. IEEE Computer Society, 2014.

[23] J. D. Kelleher, B. M. Namee, and A. DArcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press, 2015.

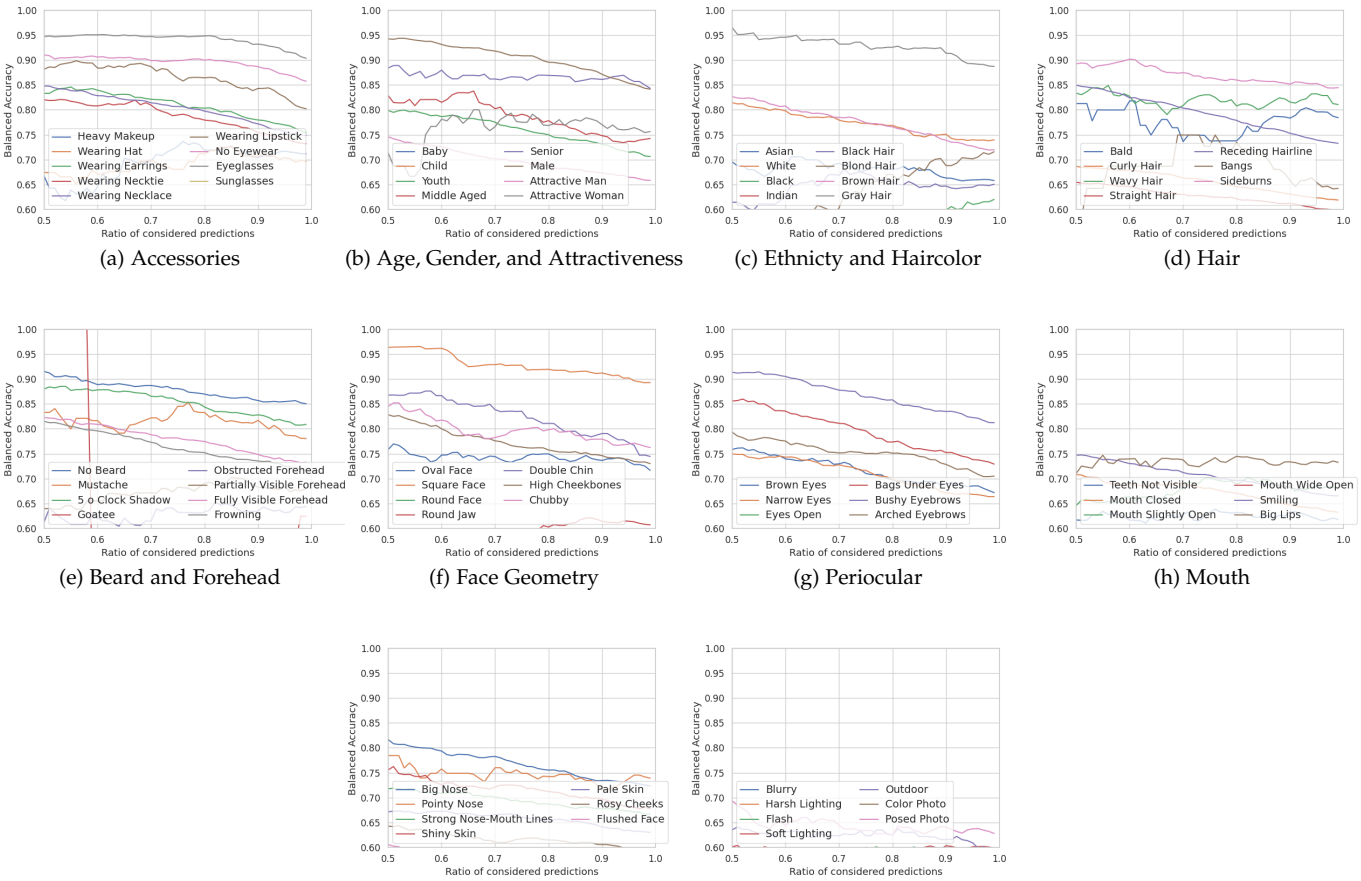
[24] E. J. Kindt. *Biometric Data, Data Protection and the Right to Privacy*. Springer Netherlands, Dordrecht, 2013.

[25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[26] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 365–372. IEEE Computer Society, 2009.

[27] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describ-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2021.3093920, IEEE Transactions on Biometrics, Behavior, and Identity Science



(a) Accessories (b) Age, Gender, and Attractiveness (c) Ethnicity and Haircolor (d) Hair

(e) Beard and Forehead (f) Face Geometry (g) Periocular (h) Mouth

(i) Nose and Skin (j) Others

Fig. 9: Accuracy-Reliability plots for the LFW database on ArcFace embeddings. The balanced accuracy of the MAC is shown for continuous RCP range of [0.5, 1]. The MAC performance of the 73 attributes is divided into 10 categories represented by subfigures (a)-(j) to allow a simple category-based analysis.

able visual attributes for face verification and image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(10):1962–1977, 2011.

[28] J. Liang, Y. Cao, C. Zhang, S. Chang, K. Bai, and Z. Xu. Additive adversarial learning for unbiased authentication. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[29] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[30] V. Mirjalili, S. Raschka, and A. Ross. Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. *IEEE Access*, 7:99735–99745, 2019.

[31] V. Mirjalili, S. Raschka, and A. Ross. Privacynet: Semi-adversarial networks for multi-attribute face privacy, 2020.

[32] V. Mirjalili and A. Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*, pages 564–573. IEEE, 2017.

[33] K. P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[34] A. A. Othman and A. Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In L. Agapito, M. M. Bronstein, and C. Rother, editors, *Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II*, volume 8926 of *Lecture Notes in Computer Science*, pages 682–696. Springer, 2014.

[35] A. J. OToole, C. D. Castillo, C. J. Parde, M. Q. Hill, and R. Chellappa. Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, 22(9):794 – 809, 2018.

[36] G. Özbulak, Y. Aytar, and H. K. Ekenel. How transferable are cnn-based features for age and gender classification? In A. Brömme, C. Busch, C. Rathgeb, and A. Uhl, editors, *2016 International Conference of the Biometrics Special Interest Group, BIOSIG 2016, Darmstadt, Germany, September 21-23, 2016*, volume P-260 of *LNI*, pages 39–50. GI / IEEE, 2016.

[37] C. J. Parde, C. D. Castillo, M. Q. Hill, Y. I. Colon, S. Sankaranarayanan, J. Chen, and A. J. O'Toole. Face and image representation in deep CNN features. In *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*, pages 673–680. IEEE Computer Society, 2017.

[38] C. J. Parde, Y. Hu, C. D. Castillo, S. Sankaranarayanan, and A. J. O'Toole. Social trait information in deep convolutional neural networks trained for face identification. *Cognitive Science*, 43(6), 2019.

[39] P. Samangouei and R. Chellappa. Convolutional neural networks for attribute-based active authentication on mobile devices. In *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*, pages 1–8. IEEE, 2016.

[40] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society, 2015.

[41] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, Jan. 2014.

[42] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper. Suppressing gender and age in face templates using incremental variable elimination. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*. IEEE, 2019.

[43] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper. Beyond identity: What information is stored in biometric face templates? In *2020 International Joint Conference on Biometrics, IJCB 2020, Houston, USA, Sept. 28 - Oct. 1, 2020*. IEEE, 2020.

[44] P. Terhörst, D. Fährmann, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Maad-face: A massively annotated attribute dataset for face images. *CoRR*, abs/2012.01030, 2020.

[45] P. Terhörst, M. Huber, N. Damer, F. Kirchbuchner, and A. Kuijper. Unsupervised enhancement of soft-biometric privacy with negative face recognition. *CoRR*, abs/2002.09181, 2020.

[46] P. Terhörst, M. Huber, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Multi-algorithmic fusion for reliable age and gender estimation from face images. In *22th International Conference on Information Fusion, FUSION 2019, Ottawa, ON, Canada, July 2-5, 2019*, pages 1–8. IEEE, 2019.

[47] P. Terhörst, M. Huber, J. N. Kolf, I. Zelch, N. Damer, F. Kirchbuchner, and A. Kuijper. Reliable age and gender estimation from face images: Stating the confidence of model predictions. In *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, Florida, USA, September 23-26, 2019*. IEEE, 2019.

[48] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognit. Lett.*, 140:332–338, 2020.

[49] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5650–5659. IEEE, 2020.

[50] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. Morales, J. Fiérrez, and A. Kuijper. A comprehensive study on face recognition biases beyond demographics. *CoRR*, abs/2103.01592, 2021.

[51] P. Terhörst, K. Riehl, N. Damer, P. Rot, B. Bortolato, F. Kirchbuchner, V. Struc, and A. Kuijper. PE-MIU: A training-free privacy-enhancing face recognition approach based on minimum information units. *IEEE Access*, 8:93635–93647, 2020.

[52] P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *8th International Workshop on Biometrics and Forensics, IWBF 2020, Porto, Portugal, April 29-30, 2020*, pages 1–6. IEEE, 2020.

[53] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274. IEEE Computer Society, 2018.

[54] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.

[55] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Feature transfer learning for face recognition with under-represented data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5704–5713. Computer Vision Foundation / IEEE, 2019.

[56] Y. Zhong, J. Sullivan, and H. Li. Face attribute prediction using off-the-shelf CNN features. In *International Conference on Biometrics, ICB 2016, Halmstad, Sweden, June 13-16, 2016*, pages 1–7. IEEE, 2016.

[57] Y. Zhong, J. Sullivan, and H. Li. Leveraging mid-level deep representations for predicting face attributes in the wild. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pages 3239–3243. IEEE, 2016.

Philipp Terhörst is a researcher at the competence center Smart Living & Biometric Technologies at Fraunhofer IGD. He received his Master of Science degree in physics in 2017 and completed his PhD in computer science from the Technical University of Darmstadt in 2021. His areas of specialization include topics in machine learning as well as biometric face recognition with a focus on quality assessment, privacy, and fairness. Dr. Terhörst is author of several publications in conferences and journals such as CVPR and IEEE Access and regularly works as a reviewer for e.g. TPAMI, TIP, PR, BTAS, ICB. For his scientific work, he received several awards, such as the 2020 EAB Biometrics Industry Award from the European Association for Biometrics for his dissertation or the IJCB 2020 Qualcomm PC Chairs Choice Best Student Paper Award. He furthermore participated in the Software Campus Program, a management program of the German Federal Ministry of Education and Research (BMBF).

Daniel Fährmann studied Applied Computer Science at the Baden-Württemberg Cooperative State University, graduating with a Bachelor of Science degree in 2013. During this period, he was employed by Hewlett-Packard GmbH as part of his dual study program. From 2012 to 2015, Mr. Fährmann continued his work at HPs as an IT consultant for VoIP and network technologies. This three-year professional experience was followed by studies in computer science at the Technical University of Darmstadt, which Mr. Fährmann completed in 2019 with a Master of Science degree. From the end of 2016 to the beginning of 2020, he worked as a research assistant in the Smart Living & Biometric Technologies department of the Fraunhofer Institute for Computer Graphics Research IGD in Darmstadt, where he has been a research associate since February 2020.

Naser Damer is a senior researcher at the competence center Smart Living & Biometric Technologies, Fraunhofer IGD. He received his master of science degree in electrical engineering from the Technical University of Kaiserslautern (2010) and his PhD in computer science from the Technical University of Darmstadt (2018). He is a researcher at Fraunhofer IGD since 2011 performing research management, applied research, scientific consulting, and system evaluation. His main research interests lies in the fields of biometrics, machine learning and information fusion. He published more than 80 scientific papers in these fields. Dr. Damer is a Principal Investigator at the National Research Center for Applied Cybersecurity ATHENE in Darmstadt, Germany. He lectures on Biometric recognition and security, as well as on Ambient Intelligence at the Technical University of Darmstadt. Dr. Damer is a member of the organizing teams of a number of conferences, workshops, and special sessions, including being a program co-chair of BIOSIG and a publication co-chair of IWBF2020. He serves as a reviewer for a number of journals and conferences and as an associate editor for the Visual Computer journal. He represents the German Institute for Standardization (DIN) in the ISO/IEC SC37 international biometrics standardization committee.

Florian Kirchbuchner is trained as an information and telecommunication systems technician and worked as an IT and telecommunications expert from 2001 to 2009 for the German Armed Forces. Afterwards he studied computer science with psychology as an application subject at the Darmstadt University of Technology and graduated with a Master of Science degree in 2014. Since then he has been working at the Fraunhofer Institute for Computer Graphics Research IGD in Darmstadt. Mr. Kirchbuchner participated in the Software Campus, a management program of the German Federal Ministry of Education and Research (BMBF) and is currently doing his PhD at the University of Technology on the topic "Electric Field Sensing for Smart Support Systems: Applications and Implications". Since 2018 he has been Head of the Department for Smart Living & Biometric Technologies at the IGD. In addition, since 2019, Mr. Kirchbuchner has been the spokesperson for the Fraunhofer Alliance Ambient Assisted Living AAL, an interdisciplinary alliance of nine institutes of the Fraunhofer-Gesellschaft for research, development and evaluation of technologies and services regarding care and homecare, prevention, therapy and rehabilitation. Florian Kirchbuchner is (co-)author of numerous scientific publications in journals, reference books and conference papers. In his role as a Principal Investigator at the National Research Center for Applied Cybersecurity ATHENE, he deals with questions concerning security, reliability and privacy of algorithms in biometric applications.

Arjan Kuijper received the M.Sc. degree in applied mathematics from Twente University, The Netherlands, the Ph.D. degree from Utrecht University, The Netherlands, and the Habitation degree from TU Graz, Austria. He was an Assistant Research Professor with the IT University of Copenhagen, Denmark, and a Senior Researcher with RICAM, Linz, Austria. His research interests include all aspects of mathematics-based methods for computer vision, graphics, imaging, pattern recognition, interaction, and visualization. He is a member of the management of Fraunhofer IGD, where he is responsible for scientific dissemination. He holds the Chair in mathematical and applied visual computing with TU Darmstadt. He is the author of over 300 peer-reviewed publications,the Associate Editor of CVIU, PR, and TVCJ, the Secretary of the International Association for Pattern Recognition (IAPR), and serves both as a Reviewer for many journals and conferences, and as a program committee member and organizer of conferences.